A ReRAM-Based Non-Volatile Flip-Flop with Sub-V_{T} Read and CMOS Voltage-Compatible Write

Ibrahim Kazi, Pascal Meinerzhagen, Pierre-Emmanuel Gaillardon, Davide Sacchetto,
Andreas Burg, and Giovanni De Micheli
EPFL, Lausanne, VD, 1015 Switzerland; Email: firstname.lastname@epfl.ch

***Abstract*— The total power budget of *Ultra-Low Power* (ULP) VLSI *Systems-on-Chip* (SoCs) is often dominated by the leakage power of embedded memories and pipeline registers, which typically cannot be power-gated during sleep periods as they need to retain data and program state, respectively. On the one hand, supply voltage scaling down to the near-threshold (near-V_{T}) or even to the sub-threshold (sub-V_{T}) domain is a commonly used, efficient technique to reduce both leakage power and active energy dissipation. On the other hand, emerging CMOS-compatible device technologies such as *Resistive Memories* (ReRAMs) enable non-volatile, on-chip data storage and zero-leakage sleep periods. For the first time, we present a ReRAM-based non-volatile flip-flop which is optimized for sub-V_{T} operation. Writing to the ReRAM devices works with a CMOS-compatible supply voltage. Thanks to near-V_{T} and sub-V_{T} operation and as compared to the write energy, which depends on the ReRAM technology, the read consumes only 5.4% of the total read+write energy. Monte Carlo simulations accounting for parametric variations in both the MOS transistors and the ReRAM devices confirm reliable data restore operation from the ReRAM devices at a sub-V_{T} voltage as low as 400 mV, and a standard deviation of up to 5% of the nominal value of the ReRAM resistance.**

I. Introduction

Ultra-Low Power (ULP) VLSI systems such as wireless sensor nodes [1] and biomedical implants [2] running for many days or even for several years have extremely low power budgets. Embedded memories and pipeline registers consume a dominant share of the total power and area of such systems [3], while for the digital signal processing core the share is often small. This power dominance is further exacerbated for systems with only short active computational periods and long sleep periods requiring data and program state retention. In this case, the leakage power of embedded memories and registers accounts for almost all the power consumption.

While near-threshold (near-V_{T}) and sub-threshold (sub-V_{T}) operation enables extremely low leakage power, emerging device technologies allowing the integration of non-volatile memory devices on CMOS chips bear the potential of zero-leakage sleep states [4]. Among many technological options, *Oxide Memories* (OxRAMs) [5] are a promising candidate for next generation, CMOS-compatible, non-volatile memory applications. Compared to traditional Flash memories, OxRAMs have better scalability and faster programming time. While a lot of research effort targets OxRAM-based standalone memories, this work focuses on the seamless integration of OxRAM devices into CMOS flip-flops for use as non-volatile, distributed storage elements.

Previous work on non-volatile flip-flops was based on the "memristor" [6], [4], on bipolar OxRAM [7], and *Magnetic Tunneling Junction* (MTJ) devices [8], [9], [10]. All this work considered circuit operation at a high supply voltage, normally corresponding to the CMOS technology's nominal voltage.

In this paper, we design a non-volatile flip-flop exploring the benefits of OxRAM devices. We combine for the first time the advantages of sub-V_{T} and near-V_{T} circuit operation with OxRAMs thereby enabling VLSI systems with ultra-low active energy dissipation in addition to non-volatile memory storage with zero-leakage. The ULP and especially the biomedical design community often prefers to use mature technology nodes for 1) high reliability; 2) low leakage currents; and 3) low cost. Therefore, this study adopts a mature 0.18 µm CMOS process. The proposed non-volatile flip-flop, to be used in pipeline registers or within standard-cell based memories [11], operates reliably in the sub-V_{T} regime. Indeed it reliably recovers the saved data on wake-up with a sub-V_{T} supply voltage and a standard deviation of up to 5% of the nominal value of the ReRAM resistance. In the proposed design, write energy is ReRAM technology dependent while the read energy can be optimized at circuit level. Thanks to the sub-V_{T} operation, the read energy has been drastically reduced down to 5.4% of the total read+write energy. Beside the main novelty of designing hybrid CMOS/OxRAM circuits for reliable operation in the sub-V_{T} and near-V_{T} regime, a number of additional factors distinguishes this work from previous work: 1) All simulations are based on real CMOS technology data (while some previous work used predictive technology models); 2) the OxRAM devices have been fabricated, characterized, and modeled in-house; 3) parametric variations are considered not only for the MOS transistors, but also for the OxRAM devices; 4) energy characterization has been done for read and write operations.

The remainder of this paper is organized as follows. Section II introduces the manufactured ReRAM stacks. Section III discusses the proposed architecture. Simulation results are given in Section IV and the paper is concluded in Section V.

II. Resistive Memory: Manufacturing Process and Switching Characteristics

Among many ReRAM candidates, OxRAMs base their working principle on the change in resistance of an oxide layer. Different physical mechanisms can be identified in the switching of ReRAMs [5]. In the following, we will focus only on the *Bipolar Resistive Switching* (BRS) [12], related to the O_2 vacancy redistribution in TiO_2 layers upon application of a voltage across the oxide. We realized memory stack prototypes of $Al/TiO_2/Al$ from bulk-Si wafers passivated by a 100-nm thick Al_2O_3 layer. 70-nm thick *Bottom Electrode* (BE) lines were patterned by lift-off and e-beam evaporation. Then, a 50-nm thick TiO_2 layer was deposited by *Atomic Layer Deposition* (ALD) at 200°C. Finally, vertical *Top Electrode* (TE) lines were defined with a second lift-off step together with contact areas used for electrical characterization. Such nodes are expected to be embedded within standard top-layer metal vias.

As opposed to most ReRAMs, our devices do not require a forming operation. Instead, the resistive switching functionality is obtained by cycling the memory. After 50 cycles,

the resistive switching behavior stabilizes to the behavior shown in Fig. 1. Consistent BRS with a *High Resistance State* (HRS) and a *Low Resistance State* (LRS) is achieved. The SET and RESET threshold voltages range from $-2\,\mathrm{V}$ to $+2\,\mathrm{V}$. Moreover, the switching operation is limited by a low current compliance of $10\,\mu\mathrm{A}$, allowing the use of small (close to minimum size) programming transistors. As opposed to this, most previously reported ReRAMs require much higher current (around 1–10 mA) to switch successfully, which needs prohibitively wide transistors to drop a sufficiently high voltage across the ReRAM.

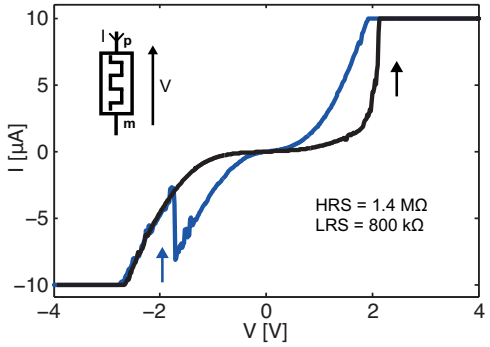

Fig. 1. $1.5\,\mu\mathrm{m}^2$ $\mathrm{Al/TiO_2/Al}$ ReRAM stack switching under $10\,\mu\mathrm{A}$ current compliance.

III. Non-Volatile Flip Flop Architecture and Operation

This section explains the design and the operating principle of the proposed ReRAM-based non-volatile flip-flop. A first design is suitable for operation at nominal and near-V_T supply, while a second version is specifically optimized for robust operation in the sub-V_T domain.

A. Architecture

A conventional master-slave flip-flop based on tri-state inverters serves as a starting point, as shown in Fig. 2 in blue color. In order to add non-volatility to this basic CMOS flip-flop, two ReRAM devices are inserted in the current sink of the cross-coupled inverter pair in the slave latch [13]. These ReRAM devices are used in a complementary way, i.e., one device is programmed to the HRS, while the other one is programmed to the LRS. Dedicated programming (or ReRAM write) circuits are highlighted in red color, while dedicated restore (or ReRAM read) circuits are shown in black color.

During normal operation, all ReRAM write and read circuits are disabled, and both branches of the slave latch are properly grounded through two NMOS transistors (controlled by $\overline{\mathrm{READ}}$). Consequently, the hybrid CMOS/ReRAM non-volatile flip-flop fully relies on CMOS transistors during normal operation, which are known to exhibit high endurance. The part of the circuit containing ReRAMs is only activated during the preparation of a sleep state or during wake-up. Therefore, the ReRAMs, whose endurance is not yet comparable with the one of CMOS transistors, do not switch very frequently, which guarantees high overall system endurance.

B. ReRAM Write Operation, or Flip-Flop Store

For the entire duration of ReRAM write and read, the clock needs to be silenced and kept low, as shown in Fig. 3, in order

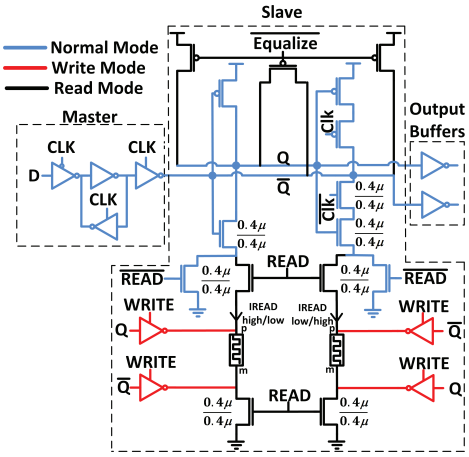

Fig. 2. ReRAM-based non-volatile flip-flop for above-V_T operation. The legend highlights the part of circuit active during different modes of operation.

for the slave latch to be non-transparent and isolated from the master. During write, the ReRAMs are disconnected from the slave latch and from the read circuits, so that the voltage drop across their terminals can be set by the write drivers (highlighted in red in Fig. 2). The write drivers are controlled by the internal nodes Q and $\overline{\mathrm{Q}}$. A write pulse width of 10 ns is used to program the ReRAMs. As illustrated in Fig. 1, a voltage of $+2\,\mathrm{V}$ or $-2\,\mathrm{V}$ is required for successful switching. To be able to use small programming transistors (with a non-negligible voltage drop across their channel) and limit the programming current, the write drivers are supplied with a voltage as high as 2.4 V. This voltage is only slightly above the nominal supply voltage range of the core transistors in the considered 0.18 µm CMOS technology and does neither seriously enhance the risk of oxide break-down, nor considerably accelerate aging.

Two architectural alternatives for the distribution of the 2.4 V supply may be adopted: 1) the supply voltage of all non-volatile flip-flops in the VLSI system is temporarily increased. This can safely be done without the need for level shifters, even if the rest of the system is biased in the sub-V_T domain, as the slave latch already holds data and the clock signal is constantly low. The energy overhead is kept small by rising only the supply of the non-volatile flip-flops; or 2) the entire VLSI system as well as the CMOS part of the flip-flop and the read circuits are constantly biased at a low supply voltage, while the CMOS write drivers are constantly supplied with 2.4 V. This alternative avoids the energy overhead associated with dynamically charging the capacitive power distribution network, but requires a level shifter in each flip-flop if the main power supply is considerably lower than 2.4 V. In this study, we adopt the first approach of dynamically rising the supply voltage during a write operation, as shown in Fig. 3.

C. ReRAM Read Operation, or Flip-Flop Restore

During system wake-up (power-on), the slave latch would ideally be directly restored, based on the data stored in the ReRAMs, during ramp-up or connection of the power supply. However, this is impossible due to a number of reasons: 1) the clock and the READ are not controlled yet; 2) there might be uncontrolled, residual charges on the internal nodes Q and $\overline{\mathrm{Q}}$; and 3) different power-gating approaches (mechanical, footer and/or header transistors, driving the supply to ground level) result in different wake-up scenarios. Therefore, the following

wake-up sequence is proposed, as shown in Fig. 3: 1) turn on the power supply; 2) at the system level, silence the clock signal to low; 3) enable the READ and the EQUALIZE; and 4) upon de-assertion of EQUALIZE, the slave latch is correctly restored based on the value of the ReRAMs. Both nodes Q and $\overline{\text{Q}}$ are pre-charged and equalized using three dedicated PMOS transistors controlled by $\overline{\text{EQUALIZE}}$.

Following this pre-charge phase, the READ is asserted. At this time, the pre-charged, internal nodes Q and $\overline{\text{Q}}$ are connected to ground through the ReRAMs. The complementary resistance state of the two ReRAMs modulates the discharge currents (the branch with HRS has a lower discharge current w.r.t. the branch with LRS), starting a race condition. As soon as one internal node is discharged to $V_{\text{DD}} - V_{\text{T,PMOS}}$, the PMOS transistor driven by that node turns on and starts to pull up the other internal node. This decides the race, before the feedback of the latch restores full logic levels.

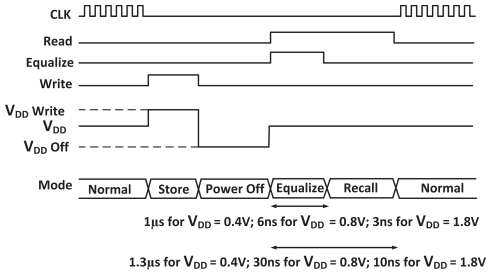

Fig. 3. Control signals sequence for ReRAM read and write operations.

D. Modifications for Robust Sub-V_{T} Operation

A correct read depends on the modulation of the discharge current by the complementary ReRAMs. However, referring to Fig. 2, the discharge current might be altered due to other reasons: 1) different pull down networks in the two branches due to the use of either a simple or a tri-state inverter; and 2) mismatch between transistor pairs (in the inverters and in the dedicated read transistors) and ReRAMs, caused by local variations. For operation in the sub-V_{T} domain (0.4 V), the following modifications are necessary to ensure correct read, as shown in Fig. 4: 1) the circuit needs to be fully symmetric; to this end, two always-on transistors (D_n and D_p) are inserted into the simple inverter to mimic the tri-state inverter, 2) all transistor pairs are upsized for better matching.

IV. Simulation Results

This section verifies the robustness, with special emphasis on the read, and characterizes the energy for the presented non-volatile flip-flops. All simulations run by Spectre assume a *Typical-Typical* (TT) process corner at 27°C. A dynamically adjustable power supply is presumed, switching between 2.4 V for write operations, and a lower value ($V_{\text{DD,read}}$) for read as well as normal operation (flip-flop sampling operation). $V_{\text{DD,read}}$ assumes the technology's nominal value (1.8 V), a near-V_{T} value (0.8 V), and a sub-V_{T} value (0.4 V). Monte Carlo circuit simulations (1000 runs) account for local parametric variations of all MOS transistors, according to statistical distributions provided by the foundry. While sophisticated statistical models of the ReRAMs are not available yet, we assume that the HRS and the LRS follow a Gaussian distribution. The measured, nominal value of HRS (1.4 MΩ) and LRS (800 kΩ) is taken as mean value, denoted by $\mu(\text{HRS})$ and

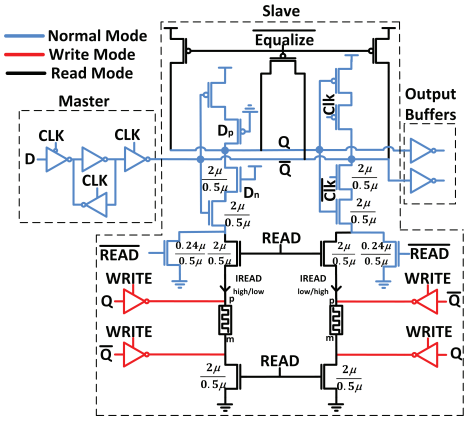

Fig. 4. ReRAM-based non-volatile optimized for robust sub-V_{T} operation. The legend highlights the part of circuit active during different modes of operation.

$\mu(\text{LRS})$. The values 40 kΩ, 80 kΩ, and 160 kΩ corresponding to 5%, 10% and 20% of $\mu(\text{LRS})$, respectively, are taken for the standard deviation, denoted by $\sigma(\text{HRS})$ and $\sigma(\text{LRS})$.

A. Sub-V_{T} Robustness Analysis

Among normal sampling, write, and read operations, read is the most critical one. Studies have shown that normal operation of CMOS flip-flops can be robust in the sub-V_{T} domain [11], while the write operation uses an elevated supply voltage. The read operation of the non-volatile flip-flop topology built for above-V_{T} operation (see Fig. 2), is simulated at 0.8 V, while the topology optimized for sub-V_{T} operation (see Fig. 4), is evaluated at both 0.8 V and 0.4 V. An appropriate metric to assess the read robustness is the initial discharge current (I_{read}) flowing through the two branches of the slave latch right after the de-assertion of EQUALIZE. Fig. 5 shows the distributions of I_{read} for the sub-V_{T}-optimized topology, at 0.4 V, for different standard deviations of HRS and LRS. For a well-controlled, repeatable ReRAM process with $\sigma(\text{HRS}) = \sigma(\text{LRS}) = 40\,\text{k}\Omega$, the discharge current flowing through the branch containing the ReRAM in the HRS is clearly lower than the current flowing through the other branch (non-overlapping I_{read} distributions). This results in zero read failures out of 1k Monte Carlo runs, as shown in Fig. 6. For a less precisely controlled ReRAM process with higher standard deviation of the resistance ($\sigma(\text{HRS}) = \sigma(\text{LRS}) = 160\,\text{k}\Omega$), the distributions of I_{read} start to overlap, which results in a small read failure probability of around 4%. Finally, Fig. 6 illustrates the high effectiveness of the proposed circuit optimizations for robust sub-V_{T} operation: the optimized circuit, supplied with 0.4 V, exhibits a much lower read failure probability than the initial, un-optimized circuit, even if the latter is supplied with a higher voltage of 0.8 V. For a badly controlled ReRAM process, rising the supply voltage of the optimized circuit from 0.4 V to 0.8 V yields a virtually zero read failure probability, while, of course, the read failure probability remains zero for a well-controlled ReRAM process.

B. Energy Characterization

Fig. 7 shows the energy dissipation of a single read and write operation of the non-volatile sub-V_{T} flip-flop (see Fig. 4). The main power supply V_{DD} (used for read and normal operations) is swept from 1.8 V to 0.4 V. Prior to a write operation the power supply is always risen to 2.4 V. For each

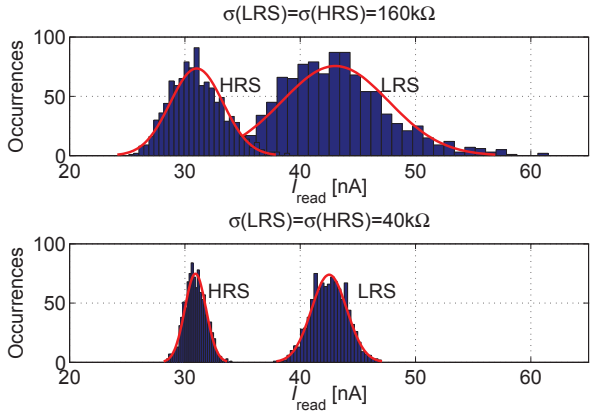

Fig. 5. Distribution of the discharge current (I_{read}) through the two branches of the slave latch of the sub-V_{T}-optimized non-volatile flip-flop, for 0.4 V, given for two different standard deviations of the ReRAM's resistance.

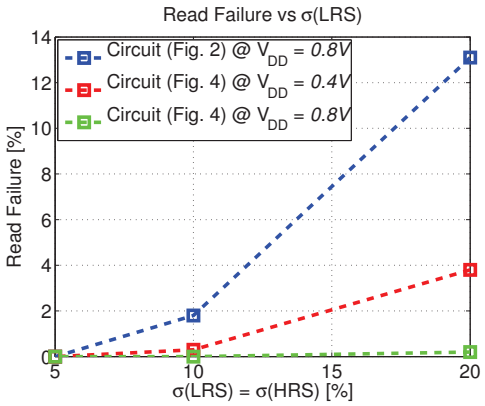

Fig. 6. Read failure probability for a ReRAM resistance's standard deviation of 5%, 10%, and 20% of the nominal LRS value. Parametric variations of MOS transistors are also accounted for, according to statistical distributions provided by the foundry.

V_{DD}, the read operation is performed at maximum speed, with the minimum required pulse widths for EQUALIZE and READ signals, given in Fig. 3. Initially, voltage scaling from 1.8 V to 0.8 V considerably reduces the read energy; however, the active energy benefits of further scaling are offset by longer pulse widths at 0.4 V (in the order of µs instead of tens of ns) and the associated integration of leakage currents.

For a main V_{DD} of 1.8 V and 0.4 V, the supply needs to be risen by 0.6 V and 2 V for a write operation, respectively. As illustrated in Fig. 7, the lower the main V_{DD} is, the larger the transition to 2.4 V, and the larger the write energy. For comparison, Fig. 7 also shows the energy cost of 5 normal sampling operations at 100 MHz, 1 MHz, and 100 kHz for 1.8 V, 0.8 V, and 0.4 V, respectively. Finally, the minimum total energy for sleep preparation and wake-up, found at 0.8 V, is 735 fJ. The write energy mostly depends on the ReRAM stack, whereas the read energy depends on the circuit topology. A direct comparison with previous work is difficult due to missing energy reports and a multitude of different ReRAMs. However, the total read+write energy of the sub-V_{T}-optimized circuit is compared with the energy of a leakage-optimized latch [14]. This shows that the proposed circuit is more energy efficient for system sleep times longer than 1.47 s.

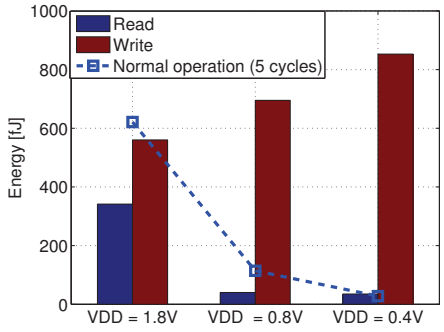

Fig. 7. Energy for read, write and five clock cycles of normal operation.

V. Conclusion

Non-volatile flip-flop circuits are designed with ReRAM technology. They leverage the use of sub-V_{T} operation enabling energy-efficient VLSI systems with zero-leakage sleep states. The manufactured oxide stacks switch their resistive state with a 0.18 µm CMOS-compatible voltage of 2 V and under a low current compliance of 10 µA. Write energy is mostly ReRAM technology dependent. Thanks to sub-V_{T} and near-V_{T} operation the read energy is brought down to 5.4% of the total read+write energy. The read energy improvement saturates between near-V_{T} and sub-V_{T} due to the increase in the READ pulse width. Monte Carlo simulations demonstrate a robust restore operation at 0.4 V, accounting for parametric variations in both ReRAM devices and MOS transistors. Robustness can be further increased by having a larger ratio between the high and low resistance values of the ReRAM.

Acknowledgment

This work was partly supported by ERC-2009-AdG-246810 and the Swiss National Science Foundation under the project number PP002-119057. P. Meinerzhagen is supported by an Intel Ph.D. fellowship.

References

[1] A. Chan *et al.*, "Low power wireless sensor node for human centered transportation system," in *IEEE SMC*, 2012.
[2] J. Abouei *et al.*, "Energy efficiency and reliability in wireless biomedical implant systems," *IEEE Tran. on Info. Tech. in Biomedicine*, vol. 15, no. 3, 2011.
[3] J. Constantin *et al.*, "TamaRISC-CS: An ultra-low-power application-specific processor for compressed sensing," in *IFIP/IEEE VLSI-SoC*, 2012.
[4] C.-M. Jung *et al.*, "Zero-sleep-leakage flip-flop circuit with conditional-storing memristor retention latch," *IEEE TNANO*, vol. 11, no. 2, 2012.
[5] G. W. Burr *et al.*, "Overview of candidate device technologies for storage-class memory," *IBM J. Res. Dev.*, vol. 52, no. 4, Jul. 2008.
[6] D. B. Strukov *et al.*, "The missing memristor found," *Nature*, vol. 453, no. 4, 2008.
[7] S. Onkaraiah *et al.*, "Bipolar ReRAM based non-volatile flip-flops for low-power architectures," in *NEWCAS*, June 2012.
[8] Y. Jung *et al.*, "An MTJ-based non-volatile flip-flop for high-performance soc," *Int. J. of Cir. Theo. and App*, 2012.
[9] W. Zhao *et al.*, "Spin-MTJ based non-volatile flip-flop," in *IEEE-NANO*, 2007.
[10] Y. Jung *et al.*, "MTJ based non-volatile flip-flop in deep submicron technology," in *ISOCC*, 2011.
[11] P. Meinerzhagen *et al.*, "Benchmarking of standard-cell based memories in the sub-VT domain in 65-nm CMOS technology," *JETCAS*, vol. 1, no. 2, 2011.
[12] Y. Chen *et al.*, "Challenges and opportunities for hfox based resistive random access memory," in *IEDM Tech. Dig*, 2011.
[13] N. Sakimura *et al.*, "Nonvolatile magnetic flip-flop for standby-power-free socs," in *IEEE CICC*, 2008.
[14] P. Meinerzhagen *et al.*, "A 500 fW/bit 14 fJ/bit-access 4kb standard-cell based sub-VT memory in 65nm CMOS," in *ESSCIRC*, Sept. 2012.